Analysis of microarray gene expression data sets

Summary

Within the last few decades, genetics has gone through major changes in scale, leading to the birth of genomics. A genomics technique of major importance is the gene expression microarray, a glass slide containing spotted probes to simultaneously measure abundance of many gene transcripts (mRNAs) within one or more samples. Computational tools and algorithms needed to be developed to analyze the data obtained, facilitated by a major increase in computational power. Chapter 1 discusses the microarray technology and analysis of the resulting data sets with respect to determination of differentially expressed genes, clustering of samples and genes, and finding genetic profiles to classify the samples.

In Chapter 2 of this thesis, focus is on classification. A novel method is described that can handle large data sets and can be tuned to specific research questions. It is applied to select gene-classifiers to discriminate mitochondrial defects (knock-outs) from other types of knock-outs in yeast. An important aspect of the method is that classifiers are prevented from being overspecific to the data set at hand. As a first step, principal component analysis (PCA) is applied, in order to select the genes that explain most of the variability within the data set. This is determined by ranking all genes based on their maximal eigenvalue-weighted contribution in the components. Use of an algorithm based on matrix decompositions prevents computation of a very large covariance matrix. Next, discriminant analysis (DA) is applied to build classifiers, in which only genes included in the limited set selected by the previous step may be incorporated. In order to allow enough genes to fulfill criteria, a non-strict cut-off is used for genes to enter the functions. However, further discriminant functions are built that distinguish between other subgroups of yeast knock-outs than the mitochondrial one. This allows selecting genes that are specifically valuable to discriminate the mitochondrial subgroup, removing highly variable genes or those specific to the particularities of the data set. This approach leads to discovery of genes related to the condition being studied. Then, classifiers are built based on those genes, using logistic modeling. Proper control of the degrees of freedom consumed prevents overspecificity of these models as well. The thiamine system appears to be a pathway of interest for the discrimination of mitochondrial yeast knock-outs. Also, models based on it perform well. The same procedure could be used to add genes from one or more other pathways, improving modeling even further.

Modeling is also the focus of the next chapters, but in this case linear modeling in order to determine which genes are differentially expressed. In Chapter 3 a modeling approach is used that allows taking into account several factors, dependent on the experimental design chosen. By incorporating features of the design directly as covariates, this modeling can also be used instead of normalizing for several factors, reducing the number of steps to take within the analysis procedure. In this chapter, a cMyBP-C knock-out mouse model is studied, for which the heterozygous mouse

shows a resembling phenotype to patients with familial hypertrophic cardiomyopathy (FHC) and the homozygous mouse develops clear hypertrophy at early age. Gene expression models are built that take labeling efficiency and, depending on the number of degrees of freedom available, sex of the animals and one or more spike (control RNA) covariates into account. Expression is compared in nine week old wild-type, heterozygous knock-out and homozygous knock-out mice. The most relevant findings are differential expression of energy related pathways, showing an increased energy demand; differential expression of the MAPK pathways, with upregulation of the JNK and p38 subsystems and downregulation of the ERK1/2 subsystem; and differential expression of apoptotis-related genes with activation of the major apoptotic route leading to caspase activity and release of cytochrome C from mitochondria. These findings are strongly linked to a pathological form of hypertrophy and are in the heterozygous mice already eminent before hypertrophy develops. Changes in cell structure and contraction related processes are present, but less prominent, possibly explained by the pre-hypertrophic (heterozygotes) or full hypertrophic state (homozygotes, where changes in gene expression may well not be detectable any more). Overall findings are similar between the two knock-out groups, with more prominent changes in the homozygous mice, which may or may not be a time effect. The usefulness of the approach used for detecting differentially expressed genes (and pathways) is clearly shown by these findings.

In Chapter 4, the value of modeling gene expression is evaluated more formally, based on Affymetrix GeneChip data sets. Two of the several commonly made assumptions in microarray data analysis are evaluated: the normality (Gaussian distribution) of the gene expression profiles and the equal dispersion of expression values for the several sample subgroups under study. The consequences of these assumptions are investigated by building several models for the expression profile of each gene and comparing their performance by the Akaike Information Criterion (AIC). This criterion corrects the minus log-likelihood (-LL) for the number of model parameters. Models are also compared using the Bayesian Information Criterion (BIC) and the unpenalized -LL, confirming results for several criteria. Also the number of non-fitting, e.g. non-converging, models is recorded. Models with and without a difference in dispersion are built for each of ten different statistical distributions: the Gaussian, Cauchy, gamma, logistic, Laplace (two parameter distributions) and the Student t, power exponential, generalized gamma, generalized logistic, skewed Laplace (three parameter distributions). Two Affymetrix data sets, considering two different kinds of human disease (cardiovascular and mitochondrial disease respectively) and using different types of chips (the hgu133a and the hgu95av2 chip respectively) are used as test data sets. The heteroscedastic Gaussian is the best fitting model amongst the ones tested. For MELAS disease, results are further evaluated, and the number of knowingly related processes retrieved increases in a heteroscedastic Gaussian approach. Also, the extra cost in computational time and degrees of freedom needed for the heteroscedastic as compared to the homoscedastic Gaussian model is very moderate. Even though the differences are

large in the test sets used, performing a test run on a subset of genes to determine the best settings and distribution for the data set at hand remains of use. In contrast, the value of the more complicated three parameter distributions appears to be very limited. Allowing for different models and different distributions for each gene separately would improve overall model fit even more, but an easy applicable method with an overall set of best settings was aimed for.

Affymetrix chips contain probesets to detect each different transcript, which consist of several probes each. Affymetrix provides information on how to group the probes. This grouping has been improved by others, because poor annotation has been considered to be the major limiting factor in quality of data analysis results for many platforms. Based on this grouping information, existing analysis methods use several approaches to calculate summarized probeset intensities from the contributing probe intensities. How this step is performed influences further results and leads to loss of between-probe variability information. As such, modeling is extended further in Chapter 5, finding out whether a multivariate normal (MVN) model directly based on the probe intensities performs better than a univariate normal (UVN) model based on summarized probeset intensities. To evaluate this modeling extension, the same data sets as in the previous chapter are used. However, in this case, it is not possible to use evaluation measures such as the AIC, BIC or -LL to find out whether a MVN model performs better than a UVN model, because the data input to the models differs. For this reason, further biologic interpretation and validation of the results is needed. Clearly fewer genes are found to be differentially expressed by the MVN modeling approach, as expected because between-probe variance is taken into account. Determining which biological pathways are affected shows that, although fewer pathways are significant because of the smaller number of differentially expressed genes, relevant changes are kept when switching to a MVN analysis. Also, the types of pathways retrieved remain the same between both analyses, and no novel ones are introduced by the MVN analysis, supporting that the last filters out non-significant findings. Furthermore, a MVN analysis leads to less non-fitting models. Where a UVN analysis already produces good results, especially in combination with a pathway approach, a MVN modeling approach still helps in focusing and performing as much a one-step and statistically sound analysis as possible.

Another design feature of Affymetrix GeneChips is the availability of a number of probesets that detect certain types of spiked-in RNA. Hybridization spikes (BioB, BioC, BioDn, CreX) are added to the sample before hybridization, while poly-A spikes (DapX, LysX, PheX, ThrX) are already added earlier, to control for amplification, labeling and hybridization. In Chapter 6 it is investigated how these spikes perform when used as covariates in both the UVN and MVN models described above. Again, the same data sets are used, where both chip types contain two probesets (five prime and three prime) for each of the hybridization spikes and three probesets (five prime, middle and three prime) for each of the poly-A spikes. For each spike, spike probeset, and – in the MVN case – probe, statistics are computed on how often they are used as

a model covariate, where the hybridization spike BioB is excluded because of its low signal intensity. It is clear that the UVN models need much more spike covariates than the MVN models. The frequency of use of each particular spike, spike probeset, and spike probe differs strongly, but results vary between the data sets used. Furthermore, the use of spikes can provide information on the experiment or the samples under consideration. It is thus advisable to incorporate all spikes, where some may later be discarded from analysis, dependent on the specific data set. The five prime probesets are more frequently used than the three prime ones, indicating a higher correspondence to sample quality and the implementation of the protocol. At the spike probe level, generally only a few (out of the twenty per probeset) make up almost the total count for their probeset. Moreover, these probes tend to be adjacent on the sequence. It may be good to limit the allowable amount of spike covariates per model, because more spike covariates appear to be more aspecifically used. However, the fact that spike covariates are used by almost all models confirms the usefulness of incorporating spikes into the experimental design.

This thesis focuses on analyzing gene expression data sets, mainly by modeling approaches. It shows that these are a good alternative to statistical testing. Modeling is more flexible and more powerful to understand the biological processes taking place. Furthermore, models do not give rise to the issues related to multiple testing, because they can be evaluated with a criterion instead of a test. Therefore, whereas tests are developed to evaluate specific hypotheses, models are especially suited for exploratory studies, such as most microarray studies. Modeling facilitates taking into account the experimental design, by incorporating several disturbing effects as covariates. Proper design is very important, including sample sizes, pairing of the samples for two-color arrays, planning of the laboratory experiments, and allowing for recording information on covariates. This thesis also shows how modeling enables a (as much as possible) one-step analysis procedure for microarray analysis, which is preferable above sequences of steps that all make assumptions and lose information. Finally, the modeling approach set-up and applied in this thesis is very general and flexible, just as the techniques used (the R package with several statistical modeling libraries), making it more adaptable to novel developments, such as proteomic arrays, exon-specific arrays, tiling arrays or even others yet to emerge. When genetic classifiers are built, it is very important to prevent overspecificity to the data set. Specific care should be taken not to exploit all degrees of freedom and ideally several different data sets must be used for building and testing the classifier. Building classifiers that reliably work for different data sets and even in several settings will become of even more importance in the application of the microarray technology for diagnosis or ‘diagnostomics’. The final chapter, Chapter 7, discusses and summarizes the contents and findings of this thesis, as well as its application to other (new) fields and possible further developments.

Samenvatting

Gedurende de laatste decennia heeft de genetica grote veranderingen in schaal doorgemaakt, wat geresulteerd heeft in het ontstaan van genomics. Een genomics techniek van groot belang is de genexpressie microarray, een glaasje waarop probes geprint zijn om tegelijkertijd de mate van aanwezigheid van vele gentranscripten (mRNAs) in een of meer monsters te meten. Computerprogrammatuur en algoritmes moesten ontwikkeld worden om de verkregen gegevens te analyseren, vergemakkelijkt door een belangrijke toename in rekenkracht. Hoofdstuk 1 bespreekt de microarray techniek en de diverse methodes om de resulterende datasets te analyseren met betrekking tot bepaling van differentieel tot expressie komende genen, clustering van monsters en genen, en het vinden van genetische profielen om monsters te classificeren.

In Hoofdstuk 2 van dit proefschrift ligt de nadruk op classificatie. Een nieuwe methode wordt beschreven die grote datasets kan verwerken en die afgestemd kan worden op specifieke onderzoeksvraagstellingen. Ze wordt toegepast om genetische classificatoren te vinden om mitochondriale defecten (knock-outs) van andere typen gist knock-outs te onderscheiden. Een belangrijk aspect van de methode is dat ze voorkomt dat classificatoren overspecifiek zijn voor de dataset onder beschouwing. Als een eerste stap wordt principal component analysis (PCA) toegepast om de genen te selecteren die de meeste variabiliteit in de dataset verklaren. Dit wordt bepaald door alle genen in de dataset te ordenen op grond van hun maximale eigenwaarde-gewogen deelname in de componenten. Het gebruik van een algoritme dat is gebaseerd op matrix decomposities, maakt dat geen enorm grote covariantiematrix berekend hoeft te worden. Hierna wordt discriminant analyse (DA) toegepast om classificatoren te bouwen, waarin alleen genen opgenomen mogen worden die in de beperkte verzameling, geselecteerd door de vorige stap, aanwezig zijn. Om voldoende genen aan de criteria te laten voldoen, wordt een niet-strikt afkappunt gebruikt voor genen om in de functies terecht te komen. Daarnaast worden echter discriminant functies gebouwd die onderscheid maken tussen andere groepen gist knock-outs dan de mitochondriale. Dit maakt het mogelijk om alleen die genen te houden die specifiek waardevol zijn om mitochondriale knock-outs van de andere te onderscheiden, waarbij hoog variabele genen of die welke gerelateerd zijn aan de eigenaardigheden van de dataset, verwijderd worden. Deze aanpak leidt tot de opsporing van genen die gerelateerd zijn aan de bestudeerde conditie. Ook worden waardevolle classificatoren gebouwd gebaseerd op deze genen, middels logistische modellering. Correcte beperking van de gebruikte vrijheidsgraden voorkomt ook voor deze modellen dat ze overspecifiek raken. Het thiamine systeem blijkt een belangrijk proces te zijn voor de discriminatie van mitochondriale gist knock-outs. Erop gebaseerde modellen presteren ook naar behoren. Dezelfde methode zou gebruikt kunnen worden om meer genen van een of meer andere processen toe te voegen, waardoor de modellering nog beter wordt.

Modellering is ook het kernpunt van de volgende hoofdstukken, maar in dit geval lineaire modellering om te bepalen welke genen differentieel tot expressie komen. In Hoofdstuk 3 wordt een modelleeraanpak gebruikt die toestaat om met verschillende factoren rekening te houden, afhankelijk van het experimenteel ontwerp dat gekozen is. Door ontwerpaspecten direct als covariaten mee te nemen, kan deze modellering ook gebruikt worden in plaats van normalisatie voor verscheidene factoren, waardoor het aantal stappen in de analyseprocedure verminderd wordt. In dit hoofdstuk wordt een cMyBP-C knock-out muizenmodel bestudeerd, waarvoor de heterozygote muis eenzelfde fenotype laat zien als patiënten met familiale hypertrofe cardiomyopathie (FHC) en de homozygote muis een duidelijke hypertrofie ontwikkelt op jonge leeftijd. Genexpressiemodellen worden gebouwd die labeling-efficiëntie en, afhankelijk van het aantal beschikbare vrijheidsgraden, geslacht van de dieren en een of meer spike (controle RNA) covariaten meenemen. Expressie wordt vergeleken in negen weken oude wildtype, heterozygote knock-out, en homozygote knock-out muizen. De meest relevante bevindingen zijn differentiële expressie van energie gerelateerde processen, die een verhoogde energiebehoefte laten zien; differentiële expressie van de MAPK processen, met upregulatie van de JNK en p38 subsystemen en downregulatie van het ERK1/2 subsysteem; en differentiële expressie van apoptose gerelateerde genen met activatie van de belangrijke apoptotische route die tot caspase activiteit en het vrijkomen van cytochroom C uit de mitochondria leidt. Deze bevindingen zijn sterk gekoppeld aan een pathologische vorm van hypertrofie en zijn in de heterozygote muizen al duidelijk aanwezig voordat hypertrofie optreedt. Veranderingen in celstructuur en contractie gerelateerde processen zijn aanwezig, maar minder uitgesproken, mogelijk verklaard door de prehypertrofe staat (heterozygoten) dan wel volledige hypertrofe staat (homozygoten, waar veranderingen in genexpressie niet meer detecteerbaar hoeven te zijn). Over het geheel genomen zijn de bevindingen dezelfde in de twee knock-out groepen, met meer prominente veranderingen in de hypertrofe homozygote muizen, wat al dan niet een tijdseffect kan zijn. Het nut van de gebruikte aanpak voor de opsporing van differentieel tot expressie komende genen (en processen) wordt duidelijk aangetoond door deze bevindingen.

In Hoofdstuk 4 wordt de waarde van het modelleren van genexpressie meer formeel geëvalueerd, gebaseerd op Affymetrix GeneChip datasets. Twee van de verscheidene geregeld gedane aannames in microarray data analyse worden geëvalueerd: de normaliteit (Gaussiaanse verdeling) van de genexpressieprofielen en de gelijke dispersie van expressiewaardes voor de verschillende groepen monsters die bestudeerd worden. De gevolgen van deze aannames worden onderzocht door verschillende modellen te bouwen voor het expressieprofiel van elk gen en hun prestaties te vergelijken middels het Akaike Information Criterion (AIC). Dit criterium corrigeert de minus log-likelihood (-LL) voor het aantal parameters in het model. Modellen worden ook vergeleken met het Bayesian Information Criterion (BIC) en de standaard -LL, de resultaten bevestigend voor verschillende criteria. Ook het aantal niet-fittende, bijvoorbeeld niet convergerende, modellen wordt

bijgehouden. Modellen met en zonder een verschil in dispersie worden gebouwd voor elk van tien verschillenden statistische verdelingen: de Gaussiaanse, Cauchy, gamma, logistische, Laplace (verdelingen met twee parameters) en de Student t, power exponential, gegeneraliseerde gamma, gegeneraliseerde logistische, skewed Laplace (verdelingen met drie parameters). Twee Affymetrix datasets, die twee verschillende types humane ziektebeelden betreffen (respectievelijk cardiovasculair en mitochondriaal) en die gebruik maken van verschillende types chips (respectievelijk de hgu133a en de hgu95av2 chip) worden gebruikt als test datasets. Het heteroscedastische Gaussiaanse model is het best passende van de geteste modellen. Voor het MELAS syndroom worden resultaten verder geëvalueerd en het aantal bekend verband houdende processen dat gevonden wordt neemt toe in een heteroscedastische Gaussiaanse aanpak. Ook zijn de extra kosten met betrekking tot rekentijd en vrijheidsgraden heel beperkt voor de heteroscedastische ten opzichte van de homoscedastische Gaussiaanse modellering. Ondanks dat de verschillen groot zijn voor de geteste datasets, blijft het van nut om voor andere datasets een testronde op een deel van de genen te doen om de beste modelleerinstellingen en de best passende verdeling te selecteren. De waarde van de meer gecompliceerde verdelingen met drie parameters blijkt daarentegen erg beperkt. Het mogelijke gebruik van een ander model en een andere verdeling voor elk gen op zich zou in totaal bekeken tot nog betere modellen leiden, maar het doel was een gemakkelijk toepasbare methode te vinden en de algehele beste instellingen.

Affymetrix chips bevatten probesets om elk verschillend transcript te detecteren, die elk uit verscheidene probes bestaan. Affymetrix voorziet in informatie hoe deze probes gegroepeerd moeten worden. Deze groepering is verbeterd door anderen, omdat slechte annotatie beschouwd is als de meest limiterende factor met betrekking tot de kwaliteit van de analyseresultaten voor veel platforms. Gebaseerd op deze groeperings-informatie, gebruiken bestaande analysemethoden verschillende technieken om samengevatte probesetintensiteiten te berekenen vanuit de bijdragende probe intensiteiten. Hoe deze stap wordt uitgevoerd, beïnvloedt verdere resultaten en leidt tot verlies van informatie over de variabiliteit tussen de probes. Daarom wordt in **Hoofdstuk 5** de modellering verder uitgebreid en wordt bekeken of een multivariaat normaal (MVN) model dat direct gebaseerd is op probe intensiteiten beter presteert dan een univariaat normaal model (UVN) gebaseerd op samengevatte probeset intensiteiten. Om deze modeluitbreiding te evalueren worden dezelfde datasets als in het vorige hoofdstuk gebruikt. Het is echter in dit geval niet mogelijk om evaluatiematen zoals de AIC, BIC of -LL te gebruiken om te beoordelen of een MVN model beter presteert dan een UVN model, omdat de gemodelleerde data verschillen. Om deze reden is verdere biologische interpretatie en validatie van de resultaten nodig. Duidelijk minder genen worden als differentieel tot expressie komend aangewezen door de MVN modellering, zoals ook verwacht omdat probe variantie meegenomen wordt. Bepaling van welke biologische processen aangedaan zijn, laat zien dat relevante veranderingen bewaard blijven bij het overgaan naar een MVN analyse, hoewel minder processen significant zijn vanwege het kleinere aantal

differentieel tot expressie komende genen. Ook de gevonden typen van processen blijven dezelfde tussen de beide analyses, en geen nieuwe worden geïntroduceerd door de MVN analyse, ondersteunend dat deze laatste niet-significante bevindingen uitfiltert. Daarnaast leidt een MVN analyse tot minder niet-passende modellen. Waar een UVN analyse al tot goede resultaten leidt, met name in combinatie met een procesaanpak, helpt een MVN aanpak nog steeds om verder te focussen en zoveel mogelijk een éénstaps en statistisch correcte analyse te verrichten.
Een ander ontwerpaspect van Affymetrix GeneChips ligt in de beschikbaarheid van een aantal probesets die bepaalde typen spiked-in RNA detecteren. Hybridisatie spikes (BioB, BioC, BioDn, CreX) worden aan het monster toegevoegd vóór hybridisatie, terwijl poly-A spikes (DapX, LysX, PheX, ThrX) al eerder toegevoegd worden om te controleren voor amplificatie, labeling en hybridisatie. In Hoofdstuk 6 wordt uitgezocht hoe deze spikes presteren als ze gebruikt worden als covariaten in zowel de UVN als MVN modellen die hierboven beschreven zijn. Opnieuw worden dezelfde datasets gebruikt, waar beide chiptypes twee probesets bevatten (vijf accent en drie accent) voor elk van de hybridisatie spikes en drie probesets (vijf accent, midden en drie accent) voor elk van de poly-A spikes. Voor elke spike, spike probeset, en – in het geval van MVN modellering – probe, worden statistieken bijgehouden met betrekking tot hoe vaak ze gebruikt worden als modelcovariaat, waarbij de hybridisatie spike BioB wordt uitgesloten vanwege haar lage signaalsterkte. Het is duidelijk dat de UVN modellen veel meer spike covariaten nodig hebben dan de MVN modellen. De gebruiksfrequentie van elke specifieke spike, spike probeset en spike probe verschil sterk, maar resultaten variëren tussen de twee gebruikte datasets. Verder kan het gebruik van spikes informatie verschaffen over het experiment of de monsters onder beschouwing. Het is dan ook aan te raden om alle spikes mee te nemen, waarna sommige dan bij de verdere analyse buiten beschouwing gelaten kunnen worden, afhankelijk van de specifieke dataset. De vijf accent probesets worden vaker gebruikt dan de drie accent probesets, indicerend dat ze meer beïnvloed worden door de monsterkwaliteit en de uitvoering van het protocol. Op het spike probe niveau, zorgen er in het algemeen slechts een paar (van de twintig per probeset) voor vrijwel het totale gebruik van de betreffende probeset. Bovendien neigen deze probes aangrenzend te zijn op de sequentie. Het is wellicht verstandig om het toegestane aantal spike covariaten per model te beperken, omdat meer spike covariaten meer aspecifiek gebruikt lijken te worden. Hoe dan ook, het feit dat spike covariaten door vrijwel al de modellen gebruikt worden, bevestigt het nut van het opnemen van spikes in het ontwerp van het experiment.
Dit proefschrift betreft analyse van genexpressie datasets, in het bijzonder door modellering. Het laat zien dat dit een goed alternatief vormt voor statistisch testen. Modellering is meer flexibel en krachtiger om de biologische processen die zich afspelen te leren begrijpen. Bovendien leidt modellering niet tot de problemen gerelateerd aan meervoudig testen, omdat modellen geëvalueerd kunnen worden met behulp van een criterium in plaats van een test. Daarom zijn modellen, terwijl testen ontwikkeld zijn om een specifieke hypothese te onderzoeken, bijzonder

geschikt voor verkennende studies, zoals de meeste microarray studies. Modellering maakt het gemakkelijker om het experimenteel ontwerp mee te nemen, door diverse verstorende effecten als covariaten op te nemen. Goed ontwerp is heel belangrijk, inclusief de monsteraantallen, het koppelen van de monsters voor tweekleuren arrays, het plannen van de laboratoriumexperimenten, en het mogelijk maken informatie over covariaten te vergaren. Daarnaast, toont dit proefschrift ook aan hoe modellering een (zoveel mogelijk) éénstaps procedure voor microarray analyse mogelijk maakt, die te verkiezen is boven opeenvolgingen van stappen die elk aannames maken en informatie verliezen. Tenslotte is de modelleeraanpak die in dit proefschrift wordt opgezet en toegepast, zeer generiek en flexibel, net als de gebruikte technieken (het R-pakket met verscheidene statistische modelleerbibliotheken), wat haar beter aanpasbaar maakt aan nieuwe ontwikkelingen zoals proteoom (eiwit) arrays, exonspecifieke arrays, tiling arrays of zelfs andere typen die nog gaan verschijnen. Wanneer genetische classificatoren gebouwd worden, is het heel belangrijk overspecificiteit voor de gebruikte dataset te voorkomen. In het bijzonder moet opgelet worden niet alle vrijheidsgraden op te gebruiken en idealiter dienen verschillende datasets gebruikt te worden om de classificator te bouwen en te testen. Het bouwen van classificatoren die betrouwbaar werken voor verschillende datasets en zelfs in verschillende settings zal nog belangrijker worden bij de toepassing van de microarray technologie voor diagnostische doeleinden, ‘diagnostomics’ te noemen. Het laatste hoofdstuk, Hoofdstuk 7, beschrijft de inhoud en bevindingen van dit proefschrift in het kort en bediscussieert ze, net zoals de toepassing ervan op andere (nieuwe) velden en mogelijke verdere ontwikkelingen.